BIOINFORMATICS

Vol. 20 no. 5 2004, pages 644–652
DOI: 10.1093/bioinformatics/btg462

Optimization models for cancer classification: extracting gene interaction information from microarray expression data

Alexey V. Antonov[1], Igor V. Tetko[1,], Michael T. Mader[1], Jan Budczies[1] and Hans W. Mewes[1,2]*

[1]GSF National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany and [2]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

Received on August 7, 2003; revised and accepted on October 6, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: Microarray data appear particularly useful to investigate mechanisms in cancer biology and represent one of the most powerful tools to uncover the genetic mechanisms causing loss of cell cycle control. Recently, several different methods to employ microarray data as a diagnostic tool in cancer classification have been proposed. These procedures take changes in the expression of particular genes into account but do not consider disruptions in certain gene interactions caused by the tumor. It is probable that some genes participating in tumor development do not change their expression level dramatically. Thus, they cannot be detected by simple classification approaches used previously. For these reasons, a classification procedure exploiting information related to changes in gene interactions is needed.

Results: We propose a MAximal MArgin Linear Programming (MAMA) method for the classification of tumor samples based on microarray data. This procedure detects groups of genes and constructs models (features) that strongly correlate with particular tumor types. The detected features include genes whose functional relations are changed for particular cancer types. The proposed method was tested on two publicly available datasets and demonstrated a prediction ability superior to previously employed classification schemes.

Availability: The MAMA system was developed using the linear programming system LINDO http://www.lindo.com. A Perl script that specifies the optimization problem for this software is available upon request from the authors.

Contact: antonov@gsf.de

INTRODUCTION

Microarray technology provides a systematic experimental access to gene regulation reflected by expression levels. The method proved its enormous potential to elucidate the nature of various biological processes within the cell and between cells at different states. Currently, applications in the classification of cancer types are of particular interest in medical diagnosis. Recent successful studies focused on acute leukemia (Getz *et al*., 2000; Golub *et al*., 1999), multiple tumor types (Ramaswamy *et al*., 2001), colon cancer (Alon *et al*., 1999) as well as breast cancer (e.g. van't Veer *et al*., 2002). However, microarray data analyses can also address issues concerning gene interactions thereby giving deeper insight into the molecular mechanics of the cell (Bornholdt, 2001; Kato-Maeda *et al*., 2001; Soinov *et al*., 2003).

Numerous algorithms have been proposed and effectively employed for cancer classification. The majority of them apply or combine known classification schemes developed and previously explored within other scientific areas (e.g. neuroscience). Some of these schemes include an initial dimension reduction. A very widespread technique is Principal Component Analysis (PCA) (Bicciato *et al*., 2003; Yeung and Ruzzo, 2001), which transforms data to reduce dimensions and, at the same time, attempts to preserve information on the data variability. Another technique, the Partial Least Squares (PLS) algorithm, transforms initial variables by maximizing cross-covariance with the target vector and was demonstrated to be superior to the PCA approach (Nguyen and Rocke, 2002). Both these methods use a complex weighted average of all genes in the initial datasets.

On the other hand, feature selection algorithms identify a subset of relevant, classifying genes. Such genes are selected according to their ability to separate different sample classes, i.e. to distinguish between tumor types or tumors from normal tissues. Previously developed feature selection algorithms consider the individual expression profile of a gene as a classification feature. Popular methods to select for genes employ the t-statistic (Tsai *et al*., 2003; Wahde *et al*., 2002) or the

*To whom correspondence should be addressed.

Wilcoxon score test (Antoniadis *et al*., 2003). In addition, the t-statistic can be used in conjunction with other methods, e.g. PLS (Nguyen and Rocke, 2002).

A similar approach used by Golub *et al*. (1999) and Ramaswamy *et al*. (2001) is based on the concept of an 'ideal' marker gene. The expression profile of such genes is a binary vector, with value 1 is for all the samples in class A and 0 for all the samples in class B (or vice versa; 'on'–'off'). The selection procedure is looking for marker genes; the genes with a profile similar to the binary expression profile. The signal-to-noise ratio measures how well the expression profile of a real gene approximates the ideal marker gene profile. The genes with the highest signal-to-noise ratio are chosen to build binary classifiers (Yeang *et al*., 2001).

In this work, we do not restrict the term 'feature' to a single gene expression profile but rather define it as (non-linear) functions integrating several of these profiles. The functions are selected to model in mathematical terms biological relationships among these genes and thus reflect functional relations among them. This definition of 'features' can be considered as a generalization of the previous model used by Golub *et al*. (1999) and Ramaswamy *et al*. (2001). The genes forming such features are presumed (and demonstrated) to be functionally related. Violation of the functional relations in the feature makes it possible to differentiate between cancer types. A selection procedure in this context tries to identify a group of genes, which form a feature that strongly correlates with an ideal marker gene. After constructing the ideal feature space based on the training set a simple algorithm such as weighted voting can be applied for tumor classification.

METHODS

In this section, we introduce a new method for the construction of gene features for the classification of multiple tumor types using microarray data. The section is organized as follows: first, we formulate in mathematical terms our concept of data transformation and ideal feature construction. Since for the gene expression data the number of response variables (i.e. samples/sample classes) is usually much smaller than the number of predictor variables (i.e. genes) it is possible to build ideal features in a number of alternative ways. Different criteria can be used depending on the procedures applied. In the second part of this section, we describe a new procedure for the feature selection that maximizes the margin of an ideal feature, i.e. the value that represents a distance of one particular tumor type from the others. Finally, we describe a classification procedure based on the ideal feature concept.

Definition of the ideal feature

Here, we introduce some conventions of notation used in this paper.

k, n, K	Counters and number of samples in the dataset, respectively
l, L	Counter and number of classes, respectively
X, x_i^k	Input dataset in matrix form
$\mathbf{x^k}$	Signature of training sample k (the kth row of matrix X)
$\mathbf{x}_i$	Expression profile of gene i (the ith column of matrix X)
α_i, β	Some real constant or variable depending on context (index represents a gene or some related vector)
$F, F(x)$	Normalized feature space and the mapping from the original to the feature space
y_k	Class of the kth sample
$f()$	A function
J_l, l	A set of indices for a group of genes and counter of such sets
e	Unity vector (all components equal to 1)
$\|.\|$	Norm

The basic idea is the following: via a nonlinear mapping $X \to F(X)$ the initial input data $\mathbf{x}^1, \ldots, \mathbf{x}^K \in R^I$ are mapped into feature space F. The purpose of such transformation is to achieve that for every two samples of the same class the scalar product in the feature space is equal to one and for every two objects from the different classes equal to zero. This can be described by the following equations

$$\begin{aligned} (F(\mathbf{x}^k)/\|F(\mathbf{x}^k)\|, F(\mathbf{x}^n)/\|F(\mathbf{x}^n)\|) &= 1, \quad \text{if } y_k = y_n \\ (F(\mathbf{x}^k)/\|F(\mathbf{x}^k)\|, F(\mathbf{x}^n)/\|F(\mathbf{x}^n)\|) &= 0, \quad \text{if } y_k \neq y_n. \end{aligned} \tag{1}$$

Let us propose one of the possible approaches to construct a mapping $F(x)$, which satisfies Equation (1). We define the ideal feature vectors as binary vectors $\mathbf{u}_l \in R^K$ with components equal to some positive scalar value z_l if the sample belongs to class l and 0 otherwise. The number of such vectors is equal to the number of tumor classes, L. Our target is to construct the ideal feature vectors in the form

$$\mathbf{u}_l = -\beta\mathbf{e} + \sum_{i \in J_l} \alpha_i f(\mathbf{x}_i), \tag{2}$$

where J_l is a set of genes that form feature l. The unity vector $-\beta\mathbf{e}$ is added to the model to include in the analysis features that can be transformed to the ideal by simple shift of every component. Thus, the vector $\sum_{i \in J_l} \alpha_i f(\mathbf{x}_i)$ has values equal to $z_l + \beta$ if the sample belongs to class l and β otherwise.

From a biological point of view the ideal feature model assumes that expression levels of genes from the set J_l are subjected to multiple positive and negative correlations with each other. The degree of the correlation remains constant in all classes except the target class l, where it is changed to a different value. This can be considered as a change in the functional relation among the genes that build the feature of class l. In other words, these genes show an interaction pattern typical for the tumor or state type.

This approach leaves freedom in the selection of functions $f(\cdot)$ and procedures to identify the corresponding gene sets J_l.

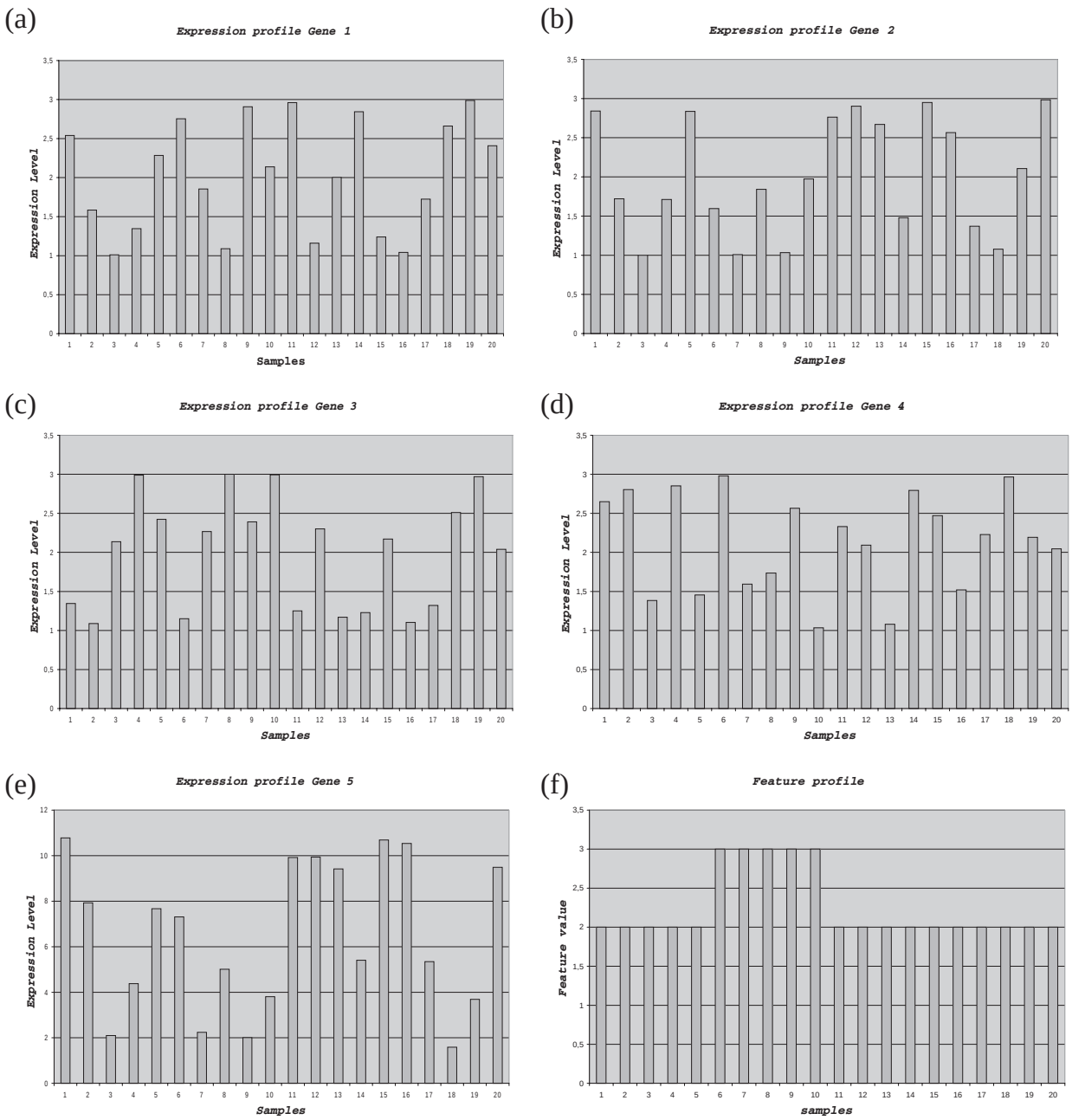

Fig. 1. The principle of ideal feature construction. (**a**), (**b**), (**c**), (**d**) and (**e**) are the log transformed expression profiles of some hypothetical genes. No single gene profile has an apparent correlation with samples 6–10, which are the members of the distinct class. However, a linear function $z = g_1 - 4g_2 + 2g_3 - g_4 + g_5$ of these profiles provides the ideal feature as shown in (**f**).

In this study, the ideal features are constructed in the form (Fig. 1)

$$\mathbf{u}_l = -\beta \mathbf{e} + \sum_{i \in J_l} \alpha_i \log(\mathbf{x}_i). \tag{3}$$

Such choice of $f(\cdot)$ implies that multiple ratios of expression levels in the corresponding group of genes remain constant. For example, if only two classes, A and B (e.g. tumor and normal tissue), are to be discriminated (in case of multiple class classification, class B includes all classes except A) one has

$$\begin{aligned} \prod_{i \in J_A} (x_i^k)^{\alpha_i} &= \mathrm{e}^{z_l + \beta}, \quad \text{for each sample } k \text{ from class A,} \\ \prod_{i \in J_A} (x_i^n)^{\alpha_i} &= \mathrm{e}^{\beta}, \quad \text{for each sample } n \text{ from class B.} \end{aligned} \tag{4}$$

Swapping of classes A and B corresponds to a renormalization of constants in Equation (3) and thus leads to the same classification result.

Optimization procedure for the ideal feature construction

For simplicity in this section we will consider only positive linear combinations in Equations (2) and (3). However, this case could be easily extended to generality by adding to the input dataset a negative copy of each gene in the form $\log(\mathbf{x}_i^-) = \mathbf{e} - \log(\mathbf{x}_i)$.

Microarray expression data tend to have a large discrepancy between the number of predictors (i.e. genes) and responses (i.e. samples). Therefore, it is possible to select classifying gene sets J_l in many different ways. Each such procedure requires some externally formulated criterion for selection among probable features. Since the number of species is much less than the number of genes, it is possible to construct a large number of ideal features.

A multiplication of coefficients α_i in Equation (4) on some constant t provides

$$\begin{aligned} \prod_{i\in J_A} (x_i^k)^{t\alpha_i} &= \mathrm{e}^{t(z_l+\beta)}, \quad \text{for each sample } k \text{ from class A,} \\ \prod_{i\in J_A} (x_i^k)^{t\alpha_i} &= \mathrm{e}^{t\beta}, \qquad \text{for each sample } n \text{ from class B,} \end{aligned} \tag{5}$$

and $t(z_l+\beta)/t\beta = (z_l+\beta)/\beta$. For this reason one can consider the ratio $(z_l+\beta)/\beta$ as margin between class A and class B and prefer features with minimal unity ($\beta\mathbf{e}$) component relative to z. If β is fixed, e.g. to 1, then this ratio is maximized by maximization of z_l. This task can be implemented using linear programming (LP). LP practically has no restrictions on the size of the problem that can be solved. Consider the following optimization problem

$$\begin{gathered} \max_{\alpha,\beta,\mathbf{s}} z_l \\ \text{subject to} \\ -\beta\mathbf{e} + \sum_{i\in J_l} \alpha_i \log(\mathbf{x}_i) + \mathbf{s} = \mathbf{u}_l; \\ |\mathbf{s}| \le \varepsilon; \;\; \alpha_i \ge 0; \;\; \mathbf{s} \in R^K. \end{gathered} \tag{6}$$

The small constant ε bounds deviations of solution of Equation (6) from the ideal feature. This constant was fixed to be 5% of $1/K$, where K is the number of training set samples.

An example for the geometric interpretation of the ideal feature generation for the two-class separation is shown in Figure 2. The constraints in optimization problem (6) ensure that class A and class B samples are lying in different parallel hyperplanes (and the distance from each sample to the corresponding hyperplane is within the constant ε) in the gene subspace formed by the gene group J_l. The value of constant β fixes the distance between hyperplane B and the origin. The objective of optimization problem (6) is to find such a gene subspace J_l, where the margin between hyperplanes A and B is maximal. The proposed procedure is referred to as MAximal MArgin linear programming (MAMA) method.

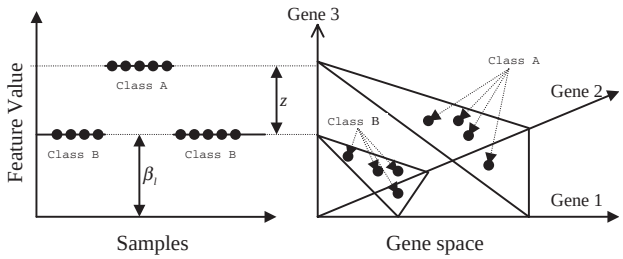

Fig. 2. Geometric interpretation of the classification procedure. The procedure is looking for a subspace of the gene space (here genes 1, 2, 3) where samples from class A and class B are lying in different parallel hyperplanes with maximum distance, z, for fixed value of β. Points represent samples.

Multiple tumor classification procedure

In case of multiple tumor classification the problem (6) is solved L times and L feature vectors $-\beta\mathbf{e} + \sum_{i\in J_l} \alpha_i \log(\mathbf{x}_i)$ are formed. The non-zero α_i elements in each solution correspond to genes that constitute group J_l, $l = 1, \ldots, L$ and are used in Equation (3). Thus, each constructed feature corresponds to one particular class (in our case tumor type). For each sample k to be classified one calculates L feature values $F_l^k = -\beta\mathbf{e} + \sum_{i\in J_l} \alpha_i \log(\mathbf{x}_i)$ for each tumor class $l = 1, \ldots, L$ and votes $P_l^k = (F_l^k - z_l - \beta)/z_l$. The index of the maximum vote l indicates the predicted class for the sample k.

RESULTS

We applied the MAMA procedure to the dataset on multiple tumor type classification (Ramaswamy *et al.*, 2001) and the dataset on acute leukemia classification (Golub *et al.*, 1999). Both datasets were downloaded from http://www.genome.wi.mit.edu/MPR

Data preprocessing and gene selection

Both analyzed datasets were subjected to prior filtering procedures. This was done to remove lowly expressed genes as well as genes invariant across samples in the training dataset. In many previous studies (Ramaswamy *et al.*, 2001), a filter based on max/min expression level ratio was applied. This measure may fail to filter lowly expressed genes or genes with only a few extreme outlier values in a few samples and stable low values in all other samples. For this reason, we applied a SD filter: the genes with the SD of expression values across the training samples less than SD threshold were filtered. Since the datasets contained some negative values, all expression values were shifted by 400 expression units. The few remaining negative values were mapped to -1

Table 1. Classification results of multiple tumor dataset (Ramaswamy *et al*., 2001) for different values of the filter threshold

Subset ID	SD threshold	Number of pre-selected genes	Misclassifications		Prediction rate (%)	
			Leave-one-out	Test samples	Leave-one-out	Test samples
1	1300	707	29	16	80	70
2	1100	905	28	16	81	70
3	1000	1042	27	14	81	74
4	900	1203	26	13	82	76
5	800	1445	25	8	83	85
6	700	1740	26	10	82	83

during the log-transform of the data matrix. The first dataset (Ramaswamy *et al*., 2001) contained about 500 duplicated gene profiles (duplicated genes had names in the dataset with and without '−2' suffix and identical expression values). The second copy (with the '−2' suffix) of these genes was removed from the dataset.

Multiple tumor type classification

The multiple classification dataset (Ramaswamy *et al*., 2001) provides measurements for 16 063 probes in 198 tumor samples representing 14 abundant human cancer classes. The dataset is split into training and test sets. The training set contains 144 samples and the test set comprises another 54 samples.

A number of different classification procedures were applied to this dataset in the original study by Ramaswamy *et al*. (2001). The best result was obtained using support vector machines—78% prediction rate on the test set (12 misclassifications of 54 test samples) and 81% prediction rate on the training set using a leave-one-out cross-validation procedure (27 misclassifications of 144 training samples). Several other classification methods analyzed yielded poor results. Their prediction rates varied from 67 to 47% depending on the applied schemes and the corresponding choice of parameters.

We tested MAMA at different filter threshold values (Table 1). The best results were calculated using a filter threshold of 800 units (subset no. 5). This filter value provided maximum prediction rates for both the leave-one-out procedure on the training set and for the prediction of the test data set. The outcome was eight misclassifications of 54 test samples (85%) and 25 misclassifications of 144 training samples (83%). The detailed analysis of results calculated for the subset 5 is shown in the Table 2.

To demonstrate that the proposed method identifies functionally related gene groups important to distinguish different tumors, the following procedure was applied. After constructing the ideal features for every tumor type, the genes involved in the feature were removed from the corresponding subset. Then new features were calculated from the remaining genes. The prediction power of these newly constructed features drops significantly. For example, for the subset 5 the prediction rate dropped from 85 to 68% (17 misclassification out of 54 samples).

Table 2. Classification results for the subset 5 from Table 1

ID	Cancer type	Training dataset			Test dataset		
		All	Correct	Misclassified	All	Correct	Misclassified
0	Breast	8	5	3	4	2	2
1	Prostate	8	5	3	6	5	1
2	Lung	8	6	2	4	4	0
3	Colorectal	8	7	1	4	4	0
4	Lymphoma	16	16	0	6	6	0
5	Bladder	8	5	3	3	1	2
6	Melanoma	8	5	3	2	2	0
7	Uterus_Adeno	8	7	1	2	2	0
8	Leukemia	24	24	0	6	6	0
9	Renal	8	5	3	3	3	0
10	Pancreas	8	6	2	3	2	1
11	Ovary	8	5	3	4	3	1
12	Mesothelioma	8	7	1	3	3	0
13	CNS	16	16	0	4	3	1
	Total	144	119	25	54	46	8

The feature profile for the CNS cancer class is shown in Figure 3, while the functional form of the feature and the genes involved are presented in Table 3. This feature mainly consists of genes exhibiting negative multiple correlations [positive α_i in the Equation (3)]. Indeed, there are only 5 out of 23 genes with negative α_i and the absolute values of them are also small compared with the positive correlations (Table 3).

Analyses of the data presented in Table 3 revealed several genes directly related to the functioning of the CNS system or/and tumor. For example, the APCL protein is a CNS-specific homologue of the adenomatous polyposis coli tumor suppressor (Nakagawa *et al*., 1998). The second gene GI O60282 (Affymetrix probe setID N98707_at) represents a neuron-specific member of kinesin family (Nagase *et al*., 1998). The involvement of genes number #4, #5, #8, #11, #13, #15, #20 (calcium/calmodulin-dependent protein kinase plays an important role in functioning of neurons) and #21 in

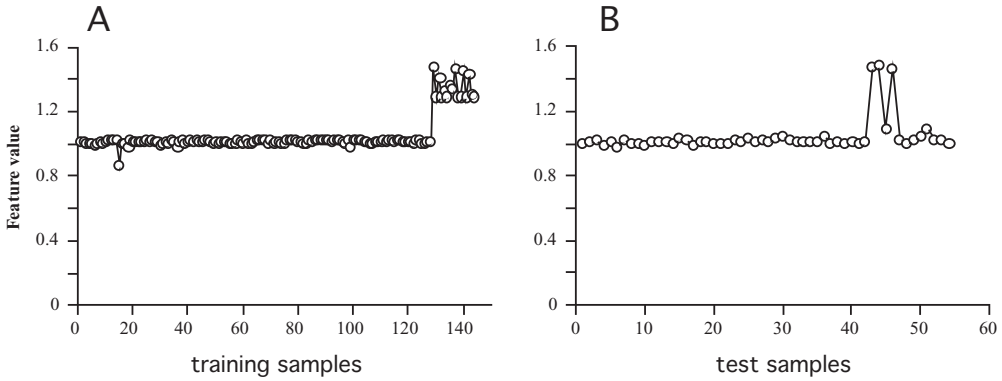

Fig. 3. Feature constructed for CNS tumor type using threshold SD = 800 (subset #5). (**A**) Training dataset. Samples 129–144 belong to the CNS tumor. (**B**) Test dataset. Samples 43–46 belong to the CNS tumor.

Table 3. Genes participating in the CNS tumor feature for subset 5 from Table 1[a]

	Probe set ID	Affymetrix gene annotation	α_j
1*	RC_D59321_f_at	*Homo sapiens* mRNA for APCL protein, complete cds	0.224
2*	N98707_at	Kinesin family member 5C	0.284
3*	RC_AA600114_at	KIAA0455 gene product	0.300
4*	C14203_s_at	EST: human fetal brain cDNA 5′ end GEN-037E11, mRNA sequence	0.287
5*	RC_AA165369_at	EST: zq49c07.s1 Stratagene hNT neuron (#937233) *H.sapiens* cDNA clone 633036 3′, mRNA sequence	0.246
6*	RC_AA284767_at	EST: zt21h07.s1 Soares ovary tumor NbHOT *H.sapiens* cDNA clone 713821 3′, mRNA sequence	0.105
7*	RC_AA478104_at	EST: zt89c03.s1 Soares testis NHT *H.sapiens* cDNA clone 729508 3′, mRNA sequence	0.146
8*	S40719_s_at	Glial fibrillary acidic protein (GFAP)	0.133
9	RC_AA598689_at	EST: ae49a08.s1 Stratagene lung carcinoma 937218 *H.sapiens* cDNA clone 950198 3′, mRNA sequence	0.116
10	RC_AA235803_i_at	EST: zs42g06.s1 Soares NhHMPu S1 *H.sapiens* cDNA clone 687898 3′, mRNA sequence	−0.092
11	M13577_at	Myelin basic protein (MBP)	0.024
12	N94832_at	EST: yy63f07.r1 *H.sapiens* cDNA clone 278245 5′	0.051
13	RC_AA007153_at	EST: 13cDNA40-3.seq Soares infant brain 1NIB *H.sapiens* cDNA clone HY18-44 3′, mRNA sequence	0.024
14	RC_AA436146_f_at	EST: zv22a12.s1 Soares NhHMPu S1 *H.sapiens* cDNA clone 754366 3′, mRNA sequence	0.015
15*	H46792_at	Fatty acid binding protein 7, brain	0.027
16	HG2815-HT2931_at	Myosin, light chain, alkali, smooth muscle (Gb:U02629), non-muscle, Alt. Splice 2	−0.012
17	RC_AA479299_at	EST: zv21f04.s1 Soares NhHMPu S1 *H.sapiens* cDNA clone 754303 3′, mRNA sequence	0.021
18	D31286_at	*H.sapiens* mRNA for smallest subunit of ubiquinol-cytochrome c reductase, complete cds	−0.006
19	U60644_at	HU-K4 mRNA	0.014
20	RC_AA398221_at	EST: zt59e10.s1 Soares testis NHT *H.sapiens* cDNA clone 726666 3′ similar to SW:KCCB_MOUSE P28652 Calcium/calmodulin-dependent protein kinase type ii beta chain; mRNA sequence	0.011
21	M15517_cds5_at	TTR gene (prealbumin) extracted from human mutant prealbumin gene directly linked to familial amyloidotic polyneuropathy (FAP)	0.011
22	D79205_at	Ribosomal protein L39	−0.005
23	M29873_s_at	Human cytochrome P450-IIB (hIIB3) mRNA, complete cds	−0.003

[a]The second column specifies Affymetrix ID and the third one indicates Affymetrix annotation for this particular gene. The last column indicates the corresponding coefficient from Equation (3). Star '*' indicates genes also found for CNR features calculated with other thresholds.

the CNS is suggested by the gene/clone annotation. The gene #7, GFAP, is involved in the differentiation of glial cells and astrocytes, a process likely to be misregulated in (undifferentiated) brain tumors. Glycoprotein m6b (gene #12, GPM6B) is a member of the myelin proteolipid protein (PLP) family and likely to be involved in neural development. Therefore, more than half of all genes present in the feature are directly related to CNS functioning. It is possible that other, not yet functionally characterized, genes in the feature are involved in oncogenesis. The annotation of a number of other genes

Table 4. Classification results for the acute leukemia dataset Golub *et al.* (1999)

ID	SD threshold	Number of pre-selected genes	Misclassifications		Prediction rate (%)	
			Leave-one-out	Test samples	Leave-one-out	Test samples
1	3000	132	2	0	95	100
2	2500	185	1	0	98	100
3	2000	273	3	0	92	100
4	1500	373	2	0	95	100
5	1000	549	2	0	95	100
6	500	1120	2	3	95	92

contains keywords such as 'tumor', 'carcinoma', etc. indicating their association with (malignant) cancer. However, we cannot exclude that some other genes may not be functionally related to CNS tumors. Since each feature differentiates particular tumor samples from all others, it can incorporate genes specific for other tumor types, e.g. prostate or ovary tumors.

Acute leukemia

The acute leukemia dataset (Golub *et al*., 1999) is one of the most intensively studied. It contains expression profiles of 7129 probe sets (i.e. genes) from 72 samples collected from acute leukemia patients. Forty-seven of these samples were diagnosed as acute lymphoblastic leukemia (ALL) and the other 25 as acute myeloid leukemia (AML). Following the experimental setup described in (Golub *et al*., 1999), the dataset has been split into a training set of 38 samples (27 ALL and 11 AML), and a test set of 34 samples (20 ALL and 14 AML).

A number of papers reported results for various procedures such as support vector machines (SVMs) (Furey *et al*., 2000), PCA (Bicciato *et al*., 2003) and partial least squares (PLS) (Nguyen and Rocke, 2002). This set can be considered as an established benchmark for any new microarray classification procedure. Prediction rates for the test set reported previously range from 86 to 97% (Bicciato *et al*., 2003; Furey *et al*., 2000; Golub *et al*., 1999; Nguyen and Rocke, 2002). On the training set using a leave-one-out cross-validation procedure some studies achieved 100% prediction accuracy (Nguyen and Rocke, 2002).

Our results with MAMA compared with the best ones reported so far. Table 4 summarizes the results for various filter thresholds. The test set prediction rate was 100% over a wide range of parameters. The best training set result, 98% of correct predictions, was received on a subset of 185 genes extracted with a SD threshold of 2500 units with the leave-one-out procedure.

In contrast to all previous studies, MAMA classified sample #66 (sample 28 of the test data) correctly. Hence, MAMA is the first method that achieved 100% prediction accuracy on the test set. The confidence of classification for this sample was not very high (Fig. 4b), nevertheless, it was correctly predicted in a wide range of parameters. The calculated feature (Fig. 4) is very close to the ideal one on the training samples and demonstrates its high prediction power on the test dataset. Even though, MAMA does not require genes in the feature to be differential between classes (see above), we found several genes involved in tumor growth (e.g. vimentin and thymosin beta 4) or in lymphocyte activation/function (e.g. lymphotoxin beta and natural killer cell transcript 4, Table 5).

DISCUSSION

The developed method is inspired by intertwining the concepts of SVM (Vapnik, 1998) and PLS (Wold, 1966). The linear SVM method, if applied to a two-class separation problem, maximizes the distance between a hyperplane and the closest samples from each class. This is done by means of the following optimization problem: $\min \|w^2\|$ subject to $y_i(wx_i + b) \geq 1$ for all i, where $y_i = \{-1, 1\}$ are the class labels. PLS, on the other hand, is a method to construct components using linear combinations of predictor variables. PLS components are constructed to maximize the sample covariance between the response values and a linear combination of predictor variables. MAMA combines features of both methods: it maximizes the margin between classes using linear combinations of predictor variables, i.e. searching among all possible gene subspaces to find the one where the margin among classes is maximal. The SVM is a quadratic problem and, as a consequence, most weights w will differ from zero. The same is true for PLS, most predictor variables with non-zero weights participate in the construction of its components. Contrary to SVM and PLS, the solutions found by MAMA represent relatively small sets of predictor variables (as a consequence of the optimization of the linear programming problem).

Despite the high overall prediction ability of MAMA (85%) for the test set of (Ramaswamy *et al*., 2001), some tumor types such as breast cancer (two out of four or 50% of correct predictions), bladder (33%) and pancreas (67%), were particularly poorly predicted. It is interesting that other previously used classification methods had also considerable difficulties to predict these tumor types. For example, the performance

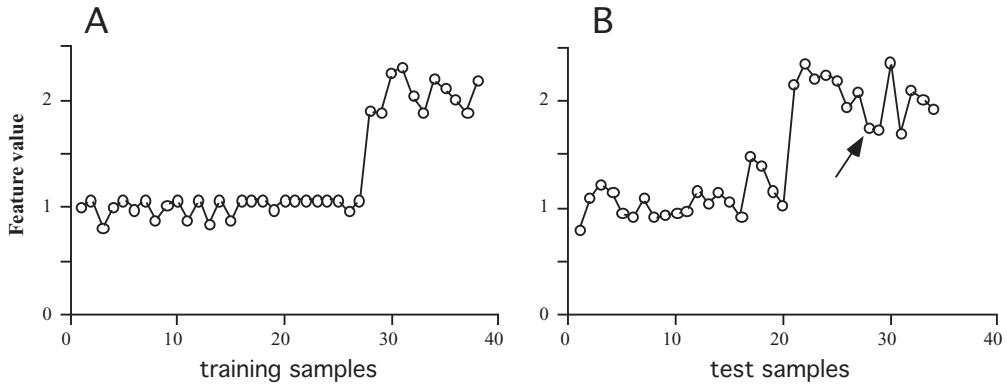

Fig. 4. Feature set constructed from the acute leukemia dataset. (**A**) Training dataset. Samples 28–38 belong to the AML class. (**B**) Test dataset. Samples 21–34 belong to the AML class. The arrow labels sample #66 that was misclassified by previous methods (Golub *et al*., 1999; Nguyen and Rocke, 2002).

Table 5. Genes participating in the classification feature from the acute leukemia dataset (subset 1)

	Probe set ID	Affymetrix gene annotation	α_j
1	M27891_at	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)	1.72
2	L38941_at	Ribosomal protein L37 (RPL37)	1.25
3	M24194_at	Alpha-tubulin mRNA	1.22
4	M19507_at	Myeloperoxidase (MPO)	0.748
5	M28130_rna1_s_at	Interleukin 8 (IL8) gene	0.498
6	U01317_cds4_at	Delta-globin gene extracted from human beta globin region on chromosome 11	0.355
7	Z19554_s_at	Vimentin (VIM)	0.245
8	M69043_at	Major histocompatibility complex enhancer-binding protein MAD3	0.202
9	U60644_at	HU-K4 mRNA	0.139
10	Y00787_s_at	Interleukin-8 precursor	0.098
11	X76223_s_at	GB DEF = MAL gene exon 4	0.032
12	X03934_at	GB DEF = T-cell antigen receptor gene T3-delta	−0.15
13	X15183_at	60S ribosomal protein L13	−0.18
14	U89922_s_at	Lymphotoxin-beta (LTB)	−0.19
15	U05259_rna1_at	MB-1 gene	−0.23
16	HG4319-HT4589_at	Ribosomal Protein L5	−0.36
17	M11722_at	Terminal transferase mRNA	−0.83

The second column specifies the Affymetrix Probe set ID and the third indicates annotation for this particular gene. The last column indicates the corresponding coefficient from Equation (3).

of SVM (Ramaswamy *et al*., 2001) was 50, 67 and 67% for breast, bladder and pancreas cancers, respectively. An algorithm developed by Bagirov *et al*. (2003) gave 25, 33 and 67% for the same tumor types. Such poor performance of classifiers is clearly inadequate to be used for treatment of patients in clinics. A simultaneous failure of various classification schemes can be attributed partially to a higher level of noise in these particular datasets. This can result from, e.g. complications with preparation of data samples for these particular tumor types. At the same time, since there is a significant variability of predictions across the aforementioned three methods (e.g. 33, 67 and 33% for bladder cancer) one can use different weighting methods to agglomerate the classifiers in order to reduce further the overall misclassification. Therefore, a use of MAMA in combination with the previously described approaches could potentially lead to the development of new powerful classification schemes.

MAMA is implemented by means of mathematical programming using the commercial package Lindo (http://www.lindo.com). This package provided a high speed of the analysis. For example, a routine analysis of 2000 genes and 150 samples required less than 5 s on a computer

Athlon 1800. Taking into account that the speed of linear mathematical programming scales approximately linearly with the number of variables (Chvatal, 1983), the LP represents an invaluable approach for microarray data analysis. The current version employs weighted voting (Hilliard, 1983) as the final classification method. However, any other classification method, especially more complex non-linear methods such as associative neural networks (Tetko, 2002) can be attached to MAMA's feature extraction procedure.

In summary, although several computational schemes have been applied successfully to multiple tumor type classification, most of them use information on single differentially expressed genes but neglect changes in gene interaction. The present work addresses this issue. A new classification scheme called MAMA was described and successfully tested on two publicly available datasets. The prediction accuracy of this procedure is high and robust in a wide range of tuning parameters. Moreover, the number of identified genes supposed to be involved in tumor development for every cancer type did not exceed 20–40.

ACKNOWLEDGEMENTS

The authors thank Stephen Rudd and Dimitrij Surmeli for their invaluable comments and suggestions. This work was supported by grant 031U118A from NGFN to W.M. INTAS 00-0363 and a BFAM grant.

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Antoniadis,A., Lambert-Lacroix,S. and Leblanc,F. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563–570.

Bagirov,A.M., Ferguson,B., Ivkovic,S., Saunders,G. and Yearwood,J. (2003) New algorithms for multi-class cancer diagnosis using tumor gene expression signatures. *Bioinformatics*, **19**, 1800–1807.

Bicciato,S., Luchini,A. and Di Bello,C. (2003) PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics*, **19**, 571–578.

Bornholdt,S. (2001) Modeling genetic networks and their evolution: a complex dynamical systems perspective. *Biol. Chem.*, **382**, 1289–1299.

Chvatal,V. (1983) *Linear Programming*. W.H. Freeman, New York.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Getz,G., Levine,E. and Domany,E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hilliard,M. (1983) *Weighted Voting Theory and Applications*. School of Operations Research and Industrial Engineering, Cornell University, p. 609.

Kato-Maeda,M., Gao,Q. and Small,P.M. (2001) Microarray analysis of pathogens and their interaction with hosts. *Cell Microbiol.*, **3**, 713–719.

Nagase,T., Ishikawa,K., Miyajima,N., Tanaka,A., Kotani,H., Nomura,N. and Ohara,O. (1998) Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*. *DNA Res.*, **5**, 31–39.

Nakagawa,H., Murata,Y., Koyama,K., Fujiyama,A., Miyoshi,Y., Monden,M., Akiyama,T. and Nakamura,Y. (1998) Identification of a brain-specific APC homologue, APCL, and its interaction with beta-catenin. *Cancer Res.*, **58**, 5176–5181.

Nguyen,D.V. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Soinov,L.A., Krestyaninova,M.A. and Brazma,A. (2003) Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol.*, **4**, R6.

Tetko,I.V. (2002) Associative neural network. *Neural Processing Lett.*, **2002**, 187–199.

Tsai,C.A., Chen,Y.J. and Chen,J.J. (2003) Testing for differentially expressed genes with microarray data. *Nucleic Acids Res.*, **31**, e52.

van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Vapnik,V.N. (1998) *Statistical Leaning Theory*. Wiley, New York.

Wahde,M., Klus,G.T., Bittner,M.L., Chen,Y. and Szallasi,Z. (2002) Assessing the significance of consistently mis-regulated genes in cancer associated gene expression matrices. *Bioinformatics*, **18**, 389–394.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiah,P.R. (ed.), *Multivariate Analysis*. Academic Press, New York, pp. 391–420.

Yeang,C.H., Ramaswamy,S., Tamayo,P., Mukherjee,S., Rifkin,R.M., Angelo,M., Reich,M., Lander,E., Mesirov,J. and Golub,T. (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17**, S316–S322.

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.